DOCUMENT RESUME

ED 135 841 95 TM 006 068

AUTHOR Kosecoff, Jacqueline; Fink, Arlene
TITLE The Appropriateness of Criterion-Referenced Tests for Evaluation Studies.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO ERIC-TM-60
PUB DATE Dec 76
CONTRACT 400-75-0015
NOTE 45p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Criteria; *Criterion Referenced Tests; Feasibility Studies; Program Effectiveness; *Program Evaluation; Test Construction; Testing Problems; *Test Reviews; Test Selection; Test Validity
IDENTIFIERS *Large Scale Evaluation

ABSTRACT

This report represents the results of an investigation of the appropriateness of criterion-referenced tests (CRTS) for large-scale evaluations. First, the development and validation of CRTS, including the formulation and generation of CRT objectives, items, and score-interpretation schemes and dimensions of item and test quality, were examined to determine whether on theoretical grounds alone CRTS are suitable for large-scale evaluations. Second, the practical characteristics of CRTS were studied to determine if it is feasible to use currently available CRTS for large-scale evaluations. A set of criteria for selecting tests for evaluation purposes was devised and used to review 28 CRTS. A conclusion was reached that CRTS are not appropriate for use in large-scale evaluations for practical but not theoretical reasons. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ERIC

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

ED135841

TM REPORT 60

DECEMBER 1976

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE APPROPRIATENESS OF CRITERION-REFERENCED TESTS FOR EVALUATION STUDIES*

Jacqueline Kosecoff and Arlene Fink

ABSTRACT

This report represents the results of an investigation of the appropriateness of criterion-referenced tests (CRTs) for large-scale evaluations. First, the development and validation of CRTs, including the formulation and generation of CRT objectives, items, and score-interpretation schemes and dimensions of item and test quality, were examined to determine whether on theoretical grounds alone CRTs are suitable for large-scale evaluations. Second, the practical characteristics of CRTs were studied to determine if it is feasible to use currently available CRTs for large-scale evaluations.

A set of criteria for selecting tests for evaluation purposes was devised and used to review 28 CRTs. A conclusion was reached that CRTs are not appropriate for use in large-scale evaluations for practical but not theoretical reasons.

INTRODUCTION

Criterion-referenced tests are becoming increasingly popular among educators and psychometricians. Perhaps the most important reason for their appearance and widespread acceptance can be traced to the new ways that had to be found to measure the effects of the educational reforms of the 1950s and 1960s. During those decades, the conventional school curriculum was declared in need of reform, and a reassessment of the goals and objectives of American education was made (19, 7, 6). Innovative courses of study and instructional technologies were subsequently developed, and programmed learning and individualized instruction became common teaching approaches. New ways of assessing student performance were needed that corresponded to these teaching innovations.

Educators have traditionally relied on paper-and-pencil achievement tests to measure learning, so it was natural for them to turn to test theoreticians to provide them with alternative ways of interpreting performance on measures of educational achievement for the new curriculums and methods of instruction. The psychometricians responded by pointing to two basic ways of assigning meaning to test scores. The first involved comparing the performance or behavior of one person or group with another person or group, and the second involved describing what a person or group can do or can be expected to do. Glaser (10) referred to these two ways of giving meaning to test scores as "norm-referenced" and "criterion-referenced," and recommended criterion-referenced score interpretations for the reformed curriculums and instruction.

The reaction to criterion-referenced tests (CRTs) was enthusiastic from the start. Because they provide score interpretations in terms of the achievement of specific and measurable skills and behaviors, CRTs have appealed to those directly responsible for the education of students and the development and evaluation of educational programs. They have also appealed to teachers who found the results of standardized tests inadequate to assist them in planning lessons and to many educators and psychologists who judged standardized, norm-referenced tests to be unfair and even biased against individuals from underprivileged and minority groups.

The interest in CRTs demonstrated by both theoreticians and practitioners has led to their frequent use for instructional diagnosis and placement and for measuring student achievement of educational tasks or objectives and professional or occupational licensure or credentials. In addition, CRTs are being suggested or used for other purposes, such as the evaluation of educational programs and the National Assessment of Educational Progress (32).

*The authors wish to thank System Development Corporation for providing support for this study.

TM006 068

The material in this publication was prepared pursuant to a contract with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Prior to publication, the manuscript was submitted to qualified professionals for critical review and determination of professional competence. This publication has met such standards. Points of view or opinions, however, do not necessarily represent the official view or opinions of either these reviewers or the National Institute of Education.

THE PURPOSES OF PROGRAM EVALUATION

The evaluation of an educational program involves the use of specific procedures that result in an appraisal of the program's merit and provides information about the nature and quality of the program's goals, outcomes, impact, and costs (9).

There are two contexts in which evaluations of educational programs are conducted. In one, an evaluation is conducted to *improve* a program, and the evaluation's clients are typically the program's organizers and staff. In the second, an evaluation is conducted to measure the *effectiveness* of a program, and the evaluation's clients are typically the program's sponsors. The context for an evaluation is determined by the information needs of the individuals and agencies who must use the evaluation information.

An evaluation is performed in an improvement context when the evaluation's clients are concerned with finding out precisely where or if a change would make the program better. Typically, the organizers of a still-developing program require this kind of information so that they can modify and improve the program. On the other hand, an evaluation is conducted in an effectiveness context when the evaluation's clients are particularly concerned with determining the consistency and efficiency with which the program achieves desired results. Those individuals who sponsored program development or who are interested in using the program require this kind of information about a well-established program's outcomes and impact.

In an effectiveness evaluation, the evaluator usually assumes a more global and independent stance toward the program than in an improvement context. In addition, the evaluator usually makes use of powerful, experimental design strategies that permit comparisons, relies on empirically validated and standardized instruments, and employs statistical and other analytic methods that allow inferences regarding the program's comparative value.

Evaluations of educational programs can be conducted for a single classroom, a grade level, a school, a district, a state, and/or for the entire nation. Large-scale evaluations encompass great numbers of students and frequently include many schools, several grades, and different districts or states.

A Study of CRTs and Large-Scale Evaluations

This report presents the results of an investigation of the appropriateness of using criterion-referenced tests for large-scale evaluations conducted in an effectiveness-evaluation context. The investigation began by examining the theory that underlies the development and validation of CRTs to determine whether, on theoretical grounds alone, CRTs are suitable or not suitable for large-scale effectiveness evaluations. The next step was to develop a set of criteria for selecting tests that are appropriate for such evaluations. Included within the set of criteria was the stipulation that the test be able to provide scores amenable to CRT interpretation. Available CRTs were then reviewed, using the identified set of criteria. Finally, conclusions were drawn based on the theoretical examination and the review.

A THEORETICAL EXAMINATION OF CRITERION-REFERENCED TESTS

A criterion-referenced test, according to Glaser and his colleagues (10), is one that is deliberately constructed to give scores that tell what kinds of behavior individuals with those scores can demonstrate. All CRTs should share several features in common:

1. They should be based on clearly defined educational tasks and purposes.
2. Test items should be specifically designed to measure the purposes and tasks.
3. Scores should be interpreted in terms of attainment of a preset criterion or level of competence with respect to the purposes and tasks.

Other definitions of CRTs have also been offered (10, 11, 13, 28). While these definitions differ considerably in terms of the limitations and constraints placed on a criterion-referenced test, they all involve reporting test scores in terms of achievement of educational tasks.

How Criterion-Referenced Tests Are Developed

To develop a CRT, test items, objectives, and score interpretations must be formulated and generated.

Formulating and generating objectives. One of the basic features of CRTs is their foundation on a clearly defined set of educational tasks and purposes. CRT objectives can be selected in at least five ways:

1. Consensus judgment. Various groups such as community representatives, curriculum experts, teachers, and/or school administrators decide which educational tasks and purposes they consider to be the most important to measure (22, 31).
2. Curriculum analysis. A team of curriculum experts analyzes a set of curriculum materials in order to identify and, where necessary, infer the educational tasks and purposes that are the focus of the test (1).
3. Expert analysis of the subject area to be tested. An in-depth analysis is made of an area, such as mathe-

Each of the eight components represents a separate section of the CRTDE rating form and is described below. (Complete copies of the CRTDE form and rating instructions are provided in the second section of this paper.) On the form, weighted items are printed in italics. Weights for classroom purposes are in parentheses and for evaluation purposes are outside the parentheses. The basic rule for applying weights is that when scores are computed by summing weights, high scores indicate better CRTs. To make the CRTDE as meaningful as possible, users of the system can choose different items for weighting or change the value of the weights.

Component 1: Marketing and Packaging

The first concern of the CRTDE is with the scope of the entire CRT across all grade levels—that is, with the content and skills it assesses and the grade or achievement levels at which forms of the CRT are available. Because program evaluations frequently involve longitudinal data collection and/or several different grade levels, CRTs that are available at many levels are particularly valuable in an evaluation context.

The next concern is with the way in which the CRT at a particular grade or achievement level is organized. A CRT's format and organization are usually determined by its intended function(s) relative to the various kinds of constraints imposed on its development and use. For example, CRTs designed as classroom aids for individualized instruction programs would have, at each grade level, many short tests each attending to a specific objective or cluster of objectives. On the other hand, CRTs designed for use in program evaluations would have fewer tests that measure more general objectives. A major feature of any CRT is that test items are designed to measure specific objectives. Consequently, it is important that the objectives be listed. The flexibility to select objectives and test items varies considerably among CRTs. Some CRTs offer a bank or pool of items each referenced to an objective from which users can create their own tests. Conversely, some CRTs offer only one pre-formatted test per grade level. Also some CRTs have two parallel or alternate forms of each test, while others do not.

Still another concern involves the materials included as standard or optional features of the CRT. The materials that are offered as part of the CRT vary considerably from publisher to publisher and can range from just a collection of tests to a system replete with audio-cassette equipment, test copies on spirit masters, and a host of resource guides. In addition to the basic CRT package, inservice training programs sometimes may be obtainable from the publisher and so may be other support services such as record keeping and computerized scoring systems.

Cost factors must also be considered when discussing the marketing and packaging of a CRT system. The cost of purchasing and administering the test must be affordable. Finally, the CRT materials must be of acceptable physical quality.

Component 2: Examinee Appropriateness

The second component of the CRTDE deals with the appropriateness of the CRT's test items, instructions, format, timing, and procedures for recording answers for examinees at the achievement or grade level designated by the publisher. In particular, the tasks, vocabulary, and level of reading required by the CRT's test items must be matched to examinees' educational experience and maturity. Similarly, instructions should be unambiguous and easily understood, and the CRT's format (the organization of printed materials on a page), illustrations and print, and auditory presentations (cassettes) must be suitable for those being tested. Finally, the timing and pacing of tests and the procedures for recording answers also must be tailored to the examinees.

Component 3: Administrative Usability

How useful are CRTs in terms of the ease with which they are administered, adapted, scored, and interpreted and their value in making educational decisions?

One factor strongly affecting a CRT's utility is the training necessary to administer the test properly. Since few schools have a staff that includes resident psychometrists, developmental psychologists, audiologists, or speech therapists, and since it is not feasible to contract for these professionals' services each time a student is tested, a CRT intended for use in a classroom context has greater utility if it can be administered by the school's regular staff and preferably, by the students' teacher, by a paraprofessional, or by the students themselves. On the other hand, this issue is not as crucial to CRTs intended for use in a program-evaluation context, since most evaluators are trained in the administration of cognitive and psychological test batteries.

Another factor closely related to test administration is the number of examinees that can be tested in a single group. In general, CRTs that have capabilities for both group and individual administration seem to be most practical. However, for individualized instruction, CRTs that can be taken individually are essential and for large-scale evaluations, CRTs that can be administered in groups are more desirable and cost effective.

The administrative usability of a CRT is also affected by the time necessary for its administration. The average attention span does not generally extend beyond 20 minutes for young children and one class period for more mature students. In addition, equipment and materials involved in test taking and the simplicity or complexity of directions can influence the ease with which a CRT is administered.

The order in which the individual tests that comprise the CRT must be administered has important consequences for a CRT's administration. For example, CRTs that require a prescribed order for testing have limited usefulness with curriculums that follow another sequence.

The ease of the scoring procedure also affects the usability of a CRT. Simple and objective hand- or machine-

scoring of tests is generally considered more desirable than difficult and subjective scoring systems. Although a CRT's usefulness may not be altered to any perceptible degree by slight variations in scoring difficulty, tests scored on a purely subjective basis are not recommended for use in large-scale evaluations.

From a pragmatic viewpoint, while ease of administration, adaptation, and scoring are desirable in a CRT, a much more basic consideration is that the scores obtained be susceptible to meaningful interpretation. The availability of interpretation guides is considered necessary to guarantee correct and consistent interpretation of CRT scores. Scoring systems of scales that are commonly used, generally understood, and that require few mathematical conversions are desirable. Similarly, scores interpretable by school staff, parents, and students are preferred to those demanding the skills of psychometrists or other specialists.

The final issue related to a CRT's administrative usability is the extent to which the test can be used to make educational decisions. Sometimes CRTs are accompanied by guidelines to translate test results into educational decisions. When used in a program evaluation context, the CRT results should permit the identification of successful and unsuccessful programs, and when used as a classroom resource, the CRT results should be able to assist teachers in assessing a student's progress and in selecting the next units of instruction. A strategy that appears to have promise in this latter regard is the referencing of objectives and test items to specific instructional materials. This strategy, often called "curriculum referencing," guides students and teachers to the appropriate materials for additional and/or supplemental instruction.

Component 4: Function and Purpose

CRTs can be used by teachers as one of their regular classroom resources in individualizing and evaluating instruction. In this classroom-management context, CRT results can be used to diagnose problems related to students' specific learning objectives; to place examinees with respect to an instructional program; to measure individuals' achievement or progress; and to assess overall learning. In an evaluation context, CRTs can be used to measure achievement, to assess the merit of an instructional program and/or to compare programs. Some CRTs are recommended by the publisher for use in a variety of contexts; others are intended for use in just one.

Component 5: Objectives Development

The issues related to the fifth component of the CRTDE include the specification of domains, the characteristics of objectives, and CRTs' match to instructional programs.

One of the basic features of CRTs is their foundation on a clearly defined set of educational tasks and purposes which together constitute the CRT's domain.* CRT objectives can be selected or defined in at least four ways:

1. *Expert judgment.* Experts assess, on the basis of their knowledge and experience in the field, the educational tasks and purposes that are the most important to measure.
2. *Consensus judgment.* Various groups such as community representatives, curriculum experts, teachers, and/or school administrators decide which educational tasks and purposes are the most important to measure (15, 22).
3. *Curriculum analysis.* A set of curriculum materials is analyzed in order to identify, and, where necessary, infer the educational tasks and purposes that should be the focus of the CRT (3).
4. *Theories of learning and instruction.* A literature review is conducted and/or consultants called in to formulate series or hierarchies of educational tasks and purposes based upon the results of psychological theory and research (13).

No matter how they are derived, educational tasks and purposes are usually called objectives or behavioral objectives. However, it should be noted that these terms have a precise meaning to educators: "An objective is an *intent* [author's italics] communicated by a statement describing a proposed change in a learner—a statement of what the learner is to be like when he has successfully completed a learning experience" (16). Developers of CRTs do not always use this definition in its purest sense. To them, an objective refers to the content that is supposed to have been learned (equivalent and nonequivalent sets in sixth-grade math, for example) and sometimes includes the behaviors the student is supposed to exhibit (naming the first five presidents of the U.S.A.).

The set of objectives or domain measured by a CRT can be characterized in terms of its organization: It can be presented without any structure, it can be organized according to major skill areas assessed by the CRT, or it can be further structured in terms of hierarchies of tasks within skill areas. Whatever organization scheme is used, it should clearly demonstrate the skeleton of the domain to be measured.

Objectives can also be characterized in terms of the rules used to write them and how broadly or narrowly they are stated. Formal rules for generating and stating objectives are needed to ensure the uniformity, manageability, and comprehensiveness of the set of objectives that the CRT measures. The level of generality at which objectives are stated is affected by the size of the domain covered by the CRT. It is possible to cover a domain by a small number of very generally stated objectives; however, objectives so stated may be ambiguous. On the other hand, detailed objectives can cover a domain in less ambiguous terms; but, to achieve this kind of clarity necessitates generating and stating a sizable and possibly unwieldy number of objectives.

Another concern closely related to domain development is the match between the CRT's objectives and those of an instructional program or curriculum. A CRT's match to a curriculum reflects the extent to which it has been designed

*The set of educational tasks and purposes that a CRT measures is sometimes called a domain or universe of content (21, 5). However, the term *domain* is used by others to mean the rules for generating test items to measure a specific objective (11). Throughout this paper, the first meaning will be used.

for use with a specific educational program (2, 20). CRTs with an extensive match to a curriculum have objectives and test items that are dependent on a particular curriculum or set of educational materials, while CRTs with some match to a curriculum, on the other hand, have objectives and test items that are only sometimes dependent on the specific tasks or purposes of an educational program. Conversely, CRTs with no match to a curriculum are based on a domain of tasks and purposes that are independent of any educational program. In a classroom context, it is generally desirable for the CRT to match the curriculum being used, while in an evaluation context, in order to be fair to all educational programs, it is usually preferred that the CRT be independent of any curriculum.

Component 6: Item Development

Once the purposes and objectives for a CRT system have been delineated, the next step is to construct and/or select tasks or test items to measure those objectives. This is one of the most difficult steps in the total test development process because there are a vast number of test items that might be constructed or generated for any given objective, even for those that have relatively narrow definitions.

Since each test item must be linked to an objective, a question arises about the number of test items that should be constructed for each objective. Some of the factors affecting the answer are the amount of testing time available and the cost of making possible interpretation errors (such as saying that a student has achieved mastery when he or she has not). More items are needed for some objectives than for others to obtain a stable estimate of learners' performance. Moreover, a set of test items that samples the range of behaviors and contents associated with an objective is more likely to give an accurate assessment of an examinee's performance than would a more restricted set of test items.

Some of the strategies and procedures used to construct test items include:

1. *Panels of experts.* A group of measurement and curriculum experts decide which items to use based on their knowledge of, and experience in, the field.
2. *A content process matrix.* Basically a variation on the classical test-construction technique, this approach involves developing for each objective a matrix of the contents and behaviors to be assessed. Items are then systematically sampled from the cells of the matrix and perhaps along a third continuum of item difficulty as well (22).
3. *Systematic item generation.* Basic "item forms" or specifications are developed for each objective that define the range of item difficulties, all the relevant contents and behaviors, and stimulus and response characteristics of items that can be used to assess the objective (10, 11, 5, 20, 19).

The procedures used to guide item writing can have a direct bearing on the utility, validity, and score interpretations of CRTs. For example, CRT systems that use specific guidelines for item construction are more likely to measure all the relevant skills and behaviors being assessed than those that do not. Moreover, specific guidelines permit the development of additional parallel test items if they are desired. Without the guidance of a systematic plan, it is very easy to construct or generate items for those aspects most amenable to measurement rather than those that might be most germane or critical. It also seems likely that responsible test developers working with an overall plan are more apt to focus their attention on the most salient (and perhaps most frequently taught) facets of an objective rather than include those components that may only be tangential to student learning. No matter what strategy is used to construct CRT items, guidelines for item writing should include comprehensive rules for the specification of tasks, conditions, and content for test items.

To what degree should items be sampled to compare their relative difficulty and possible content coverage? It is a well-known and frequently used principle of test construction that even slight changes in an item can affect its difficulty. The extent to which the items are sampled with respect to difficulty has a direct bearing on the interpretation of the scores obtained. In other words, if only the most difficult items are used to measure an objective, the phrase *achievement of the objective* will have a very different meaning than if the items are sampled over the full range of difficulties.

An issue related to item writing, and one which has perhaps not received as much attention as it should, is the potential interaction between the objective and how it is measured. It is often assumed, for example, that selected response items (such as multiple-choice questions) serve as an effective proxy for constructed response items (such as completion or short-answer questions) because the performance of students on the two kinds of items is highly related. Although this may be generally true, it may not be true for certain kinds of objectives. Further, the degree of mastery required to answer a constructed response item is usually greater than that needed to answer a selected response item. Despite the obvious advantages of the former format, the ease of scoring items using a selected response format has led to its almost exclusive use in published measures, including CRTs.

Component 7: Methods of Score Interpretation

One of the distinctive features of a CRT is its ability to provide a means for describing what an individual or group can do, knows, or feels without having to consider the skills, knowledge, or attitude of others. Consequently, CRT scores are reported and interpreted in terms of the level of performance obtained with respect to the objective(s) or domain on which the CRT is based. This type of score reporting is very different from that used for norm-referenced tests in which scores are reported in terms of the performance of other individuals or groups.

Criterion-referenced score interpretations can be expressed in several ways. CRT scores can be reported as the percentage of individuals who correctly answered each item (the item's difficulty). This score is used primarily

when only one item is tested per objective—for example, as is the case with National Assessment of Educational Progress. Reporting an individual's or group's actual level of performance as the percentage or number of items correctly answered for each objective is another very common way of expressing CRT scores. An empirical variation of this score is the estimated "true" level of performance, referring to the portion of the total universe of items for an objective that an individual or group could answer correctly. Mastery interpretation schemes report scores in terms of whether or not a pre-set performance level has been achieved, and an individual is described as having or not having mastered a given objective. Selection of the criterion level of performance for a mastery score interpretation should not be arbitrary, but should be justifiable and based on a concept of mastery. Experience, theories of learning, or experimentation can be used to justify a concept of mastery. Nonarbitrary definitions of mastery have been offered by Novick and Lewis (17), Harris (8), and others. In some of these mastery systems, several categories are employed to distinguish between degrees or levels of mastery. CRT scores also can be reported in terms of the level of performance achieved after a certain amount of learning time or as the probability of passing the next unit of instruction.

Scores on CRTs need not be limited to just a CRT interpretation. Other score interpretations can also be provided to expand upon the CRT interpretation (14, 4, 6). For example, criterion-referenced information can be combined with norm-referenced information in the following way: "This school had an average score of 5 out of 10 on the objective (a CRT interpretation) which is one standard deviation below the national average of 7 out of 10 (a norm-referenced interpretation)." The idea of using both types of score interpretations is not new and it does not reduce the theoretical soundness of the score interpretation (4, 14, 15). Combining score interpretations is useful for describing what a student can be expected to do and how exceptional or typical this performance is. Comparative or norm-referenced scores are typically reported in terms of standard score scales, age/grade equivalents, and percentiles.

The type of scheme used to report scores is, in part, determined by the context in which a CRT is used. Reporting results as the percentage of items passed per objective can be meaningful in a classroom context if the objectives are carefully matched to the curriculum. However, in an evaluation context, this type of interpretation alone may be inadequate because it provides insufficient information for decision making and loses meaning outside the classroom. For evaluation purposes, it is probably useful to supplement this score with comparative data or to use scores whose criterion levels have been validated empirically or based on theories of instruction and learning.

Component 8: Analysis and Validation

It is axiomatic that all tests and measures must be validated before basing decisions upon them. This process can involve giving the test to students and studying their responses (response data) or relying upon review by experts (judgmental data). The issues addressed in this component of the CRTDE involve both of these processes and include the characteristics of field tests conducted to certify CRTs and dimensions of item and test quality.

A CRT should be field tested on a sample of individuals who are representative of those for whom it was intended. Since most commercially published CRTs are intended for widespread use, they should be tried out in a large-scale field test with samples that are geographically and ethnically representative of the nation.

There is much ambiguity about the procedures appropriate for analyzing the data from CRTs' field tests. Nevertheless, there are several dimensions of item and test quality that are considered to be relevant to CRTs and that have associated with them review procedures, data collection strategies, experimental designs, and statistical indexes. In recognition of the uncertainty in the field with respect to the psychometric characteristics of a good CRT and the methods for measuring their presence/absence, the CRTDE system only includes the dimensions of CRT quality that are attended to by test publishers.*

There are five dimensions that can be used to assess item quality. They are:

1. *Item-objective congruence.* A test item is considered "good" if it measures or is congruent with the objective that it is supposed to assess. Item-objective congruence can be established by using judgmental data. Typically, content experts are given a variety of objectives and the item used to measure them and are asked to comment on the appropriateness of the item-objective relationship.
2. *Equivalence (internal consistency).* An item is considered "good" if it "behaves" like other items that measure the same objective. The concept is similar to item-objective congruence, but its proper use depends on response data. Equivalence is usually measured by computing the biserial correlation between the score on an item and the total score on all items measuring that objective. It should be noted that for broadly defined objectives, internal consistency will be lower than for narrowly defined objectives, and this must be taken into account when using internal consistency data to make decisions about item quality.
3. *Stability (over time).* An item is considered "good" if examinee performance is consistent from one test period to the next in the absence of any special intervention (such as instruction, which is an intervention that can change examinee performance). Stability involves response data and can be measured with the phi coefficient.
4. *Sensitivity to instruction.* An item is considered "good" if it is sensitive to instruction—that is, if the item is able to discriminate between those who have and those who have not benefited from instruction, assuming that instruction was adequate. This mea-

*Several of these dimensions—for example, item and test estimates of equivalence and stability—depend upon the variability of performance in the field test sample. To be meaningful, the sample must be representative of the population of interest and contain sufficient variation.

sure of item quality is usually computed for CRTs that are linked to particular educational programs and that require response data. Typically, examinees are tested before and after an educational program, and those items that many examinees fail before instruction but pass after instruction are considered to be sensitive to the instruction.

5. *Cultural/sex bias.* An item is considered "good" if it leads to accurate inferences about the knowledge, skills, or other attributes of an individual or group. Bias can be assessed using either judgmental or response data. If the former are used, representatives of different cultural groups, members of each sex, and/or linguists examine test items to determine whether vocabulary or content could be misinterpreted. If response data are used to assess bias, they are analyzed (typically, using ANOVA or regression) for item-cultural/sex interactions.

There are six dimensions commonly used to express the quality of a CRT. They are:

1. *Test-objective congruence.* Similar to item-objective congruence, test-objective congruence assesses the extent to which the total test or subtest measures the relevant objective. Test-objective congruence is usually determined by using judgmental data.
2. *Equivalence (internal consistency).* Test equivalence measures the homogeneity of test items for an objective—that is, how coherently the test items assess a particular objective. Equivalence can be estimated by using split-half correlations, Kuder-Richardson formulas, or coefficient alpha. It should be noted that internal consistency estimates will be lower for more broadly stated objectives.

3a. *Stability* (test-retest, or alternate forms). A test is stable to the extent that examinee responses are consistent from one test period to another or across alternate forms of a test in the absence of any intervention.

3b. *Stability* (number of items per objective). A determination is made of the number of items that should be tested in order to obtain a stable score on an objective. For this type of stability, the assumption is made that for each objective there is a pool or population of items with mixed difficulties that deals with the objective and that for any given test, a sample of those items is selected. Stability can be estimated with response data using correlation techniques and/or Bayesian models (17).

4. *Sensitivity to instruction.* This measure of test quality is usually obtained for CRTs that are linked to a specific educational program. It can be obtained from response data by comparing scores of those who have and have not received instruction.
5. *Cultural/sex bias.* Bias in measurement occurs when characteristics of the test, the testing process, or the interpretation of test results lead to inaccurate inferences about the knowledge, skills, or other attributes of individuals or groups (1). Bias can be measured by ANOVA or regression techniques using response data or by expert review using judgmental data.
6. *Criterion validity.* Criterion validity establishes the meaningfulness of the criterion in terms of which CRT scores are interpreted. Establishing criterion validity makes use of classical validity measures and is either a one-step or a two-step process:

Step 1: The first step involves assessing the meaningfulness or content validity of the domain: that objectives have been selected and organized to be in themselves educationally significant and that test items have been systematically generated to cover the objectives. Step 1 is usually established by having experts review the objectives and test items to determine the extent to which they were developed in conformance with prespecified procedures and to which they cover the domain in a comprehensive and meaningful manner.

Step 1 must be completed for all CRTs and, in some cases, is sufficient for establishing criterion validity. One example of a test that requires only Step 1 criterion validity is a CRT based on objectives that are narrowly defined and operationally stated in such detail that generating items requires only transposing the objectives into question form. CRT score interpretations of objectives with these characteristics are meaningful because the objectives describe skills that can be measured directly by test items. In a second case, the CRT's objectives are linked to a curriculum and interpreted by teachers or curriculum experts. CRT score interpretations are meaningful for these objectives because the skills and knowledge measured are those taught in classrooms using a specific curriculum. A third case in which Step 1 validity is sufficient is when comparative data are provided or when the CRT score interpretation for each objective is supplemented by a normative interpretation.

Step 2: In Step 2, criterion validity is established through empirical means and involves determining whether examinees who perform well on the test have really achieved the educational objective. Step 2 criterion validity can be measured by comparing scores obtained by individuals who, in advance of taking the CRT and using *independent* criteria, are judged to possess or not possess the skills that the objective is intended to measure. To the extent that the CRT discriminates between these two groups of individuals, the CRT has criterion validity. (Note that if an objective or domain is considered analagous to a psychological state, then Step 2 criterion validity can be likened to construct validity; otherwise, Step 2 criterion validity can be likened to concurrent validity.)

By establishing Step 2 criterion validity, the relationship between test items and the objectives they are supposed to measure is empirically confirmed. Step 2 criterion validity permits assertions about mastery of the individual objectives that comprise a domain and about more complex behaviors whose component parts are defined by the domain.

APPENDIX

RATING FORMS AND INSTRUCTIONS

The CRTDE rating system has been designed so that school personnel, test-selection committees, researchers, and professional evaluators can prepare a description and evaluation of criterion-referenced tests. In the following section of this paper, the rating form and instructions for its use are presented.

9

1: Marketing and Packaging*					
1. SCOPE (of Total CRT)	Grade Levels Tested (total CRT) *Grade Level Coverage (total CRT)*	K 1 2 3 4 5 6 7 8 9 10 11 12 0(0) ☐ *Doesn't cover needed grades*		0(1) ☐ *Covers needed grades*	
2. TEST ORGANIZATION	*Number of Separate Tests*	2(0) ☐ *1 test*	0(1) ☐ *2-9 tests*	0(2) ☐ *10 or more tests*	☐ *Not applicable. No pre-formatted tests*
	List of Objectives	☐ No	☐ Yes		
	Flexibility to Select Items	☐ No	☐ Yes, some	☐ Yes, extensive	
	Answer Sheet Format	☐ Hand scoreable only	☐ Machine scoreable only	☐ Hand or machine scoreable	
	Test Length (range)	☐ ___ to ___ objectives per test	☐ ___ to ___ items per objective		
	Alternate Forms	☐ No	☐ Yes		
3. AVAILABLE MATERIALS	Tests	☐ No	☐ Yes, standard	☐ Yes, optional	
	Technical Manual	☐ No	☐ Yes, standard	☐ Yes, optional	
	User's Manual/Guides	☐ No	☐ Yes, standard	☐ Yes, optional	
	Answer Sheets	☐ No	☐ Yes, standard	☐ Yes, optional	
	Cassettes/Special Equipment	☐ No	☐ Yes, standard	☐ Yes, optional	
	Student Report Forms	☐ No	☐ Yes, standard	☐ Yes, optional	
	Resource Books	☐ No	☐ Yes, standard	☐ Yes, optional	
4. OTHER PUBLISHER SERVICES	Inservice Training Available	☐ No	☐ Yes, standard	☐ Yes, optional	
	Scoring of Tests	☐ No	☐ Simple scores only	☐ Extensive score summaries	
5. COSTS	Cost per Student at a Given Grade Level				
6. QUALITY OF MATERIALS	Physical Quality of Materials	☐ Poor	☐ Good	☐ Very good	

*Information sources for each component of the CRTDE may be found in the CRT's test booklets, examiner's manual, and/or technical manual. It should be noted that not all CRTs have these three parts nor is information similarly organized from publisher to publisher.

2. Examinee Appropriateness				
1. TEST ITEMS	Study of Test Item's Appropriateness Vocabulary, Brevity, Clarity Tasks Required of Examinees	☐ Not reported ☐ Inappropriate ☐ Inappropriate	☐ Expert judgment	☐ Response data ☐ Appropriate ☐ Appropriate
2. INSTRUCTIONS	Vocabulary, Brevity, Clarity Illustrative Sample Items	☐ Inappropriate ☐ Not present	☐ Present but not clarifying	☐ Appropriate ☐ Effective and clarifying
3. FORMAT	Test Page Layout Illustrations and Print Auditory Presentation	☐ Complicated ☐ Unclear ☐ Garbled		☐ Clear ☐ Clear ☐ Clear
4. TIMING	Timing and Pacing	☐ Inappropriate	☐ Appropriate	☐ No guidelines
5. RECORDING ANSWERS	Response Scheme	☐ Complicated		☐ Simple

3: Administrative Usability				
1. ADMINISTRATION	*Size of Testing Group*	0(1) ☐ *Individual only*	2(1) ☐ *Groups only*	2(2) ☐ *Individual or group*
	Administrator	0(0) ☐ *Specialist*	1(1) ☐ *School personnel*	1(2) ☐ *Self-administration*
	Administration Time	☐ Shortest possible time: ____	☐ Longest possible time: ____	☐ No time limits
	Directions	☐ Not available	☐ Available, incomplete	☐ Available, complete
	Equipment	☐ Special equipment needed	☐ No special equipment needed	
	Order for Testing	☐ Prescribed sequence	☐ General guidelines	☐ None
2. ADAPTABILITY	Modification of CRT by User	☐ No	☐ Yes, limited	☐ Yes, extensive
3. SCORING	Ease of Scoring	☐ Subjective	☐ Objective	
4. INTERPRETATION	Score Interpreter	☐ Specialist	☐ School personnel	☐ Self-interpreting
	Score-Interpretation Guide	☐ Not available	☐ Available, complicated	☐ Available, not complicated
	Score Interpretability	☐ Complicated, unusual	☐ Simple, standard	
5. DECISIONS	Decision-Making Utility	☐ No guidelines	☐ Yes, limited	☐ Yes, extensive
	Curriculum Referencing	0(0) ☐ *None*	0(1) ☐ *Yes, some*	0(2) ☐ *Yes, extensive*

4: Function and Purpose				
1. PURPOSE	*Test Uses (Circle all that apply)*	0(1) ☐ *Diagnosis* 2(1) ☐ *Achievement/outcomes* ☐ Other, please specify:	0(1) ☐ *Placement* 2(0) ☐ *Comparison of instruction programs*	1(2) ☐ *Achievement/progress*

16

17

5: Objectives Development					
1. DOMAIN (OBJECTIVE) SPECIFICATION	Domain Definition Domain Structure	☐ None ☐ Content area	☐ Vague, general ☐ Content/process matrix	☐ Specific ☐ Objectives/item generation format	
2. OBJECTIVES' CHARACTERISTICS	Organization of Objectives Level of Generality of Objectives	☐ None ☐ General	☐ Simple list (no structure) ☐ Specific	☐ Categories (strands) ☐ Very detailed	☐ Hierarchy
3. MATCH TO INSTRUCTION	*Curriculum Match*	2(0) ☐ *None*	1(1) ☐ *Some*	0(2) ☐ Extensive	

6: Item Development				
1. ITEM-OBJECTIVE RELATION	*Items Coded to Objectives* Number of Items per Objective Scope of Coverage of Objectives	0(0) ☐ *No* ☐ Not applicable–Items not coded to objectives ☐ Poor	2(2) ☐ *Yes* ☐ Range: ___ to ___ items/objective ☐ Good	☐ Average: ___
2. ITEM GENERATION	Rules for Item Writing	☐ None	☐ Suggestions	☐ Specifications

20

21

7: Methods of Score Interpretation				
1. CRITERION-REFERENCED	*Scores Reported in Terms of Level of Performance (Check all that apply)*	0(0) ☐ *Item difficulty* 0(0) ☐ *Arbitrary mastery* 1(1) ☐ Achievement level after a certain amount of learning time	0(0) ☐ *Actual score (Percent of items correct)* 1(1) ☐ *Empirical mastery* 1(1) ☐ Probability of achieving next level	1(1) ☐ *True score (Percent of objective achieved)*
2. NORM-REFERENCED	*Comparative Scores Reported as well as a Criterion-Referenced Score (Check all that apply)*	1(1) ☐ *Standard score scales*	1(1) ☐ *Age/grade equivalents*	1(1) ☐ *Percentiles*

8. Analysis and Validation				
1. FIELD TEST	Field Test Reported	☐ No	☐ Yes	
	Scale	☐ Small	☐ Moderate	☐ Large
	Scope-Geographic	☐ Local	☐ Regional	☐ National
	Scope-Ethnic	☐ Little ethnic representation	☐ Ethnic representation	
	Sample Representativeness	☐ No sampling plan	☐ Probability sampling plan	☐ Special sampling (non-probability)
2. ITEM QUALITY (judgmental data)	Item-Objective Congruence	☐ Not reported	☐ Reported (Give value:)	
	Cultural Bias	☐ Not reported	☐ Reported (Give value:)	
3. ITEM QUALITY (response data)	Sensitivity to Instruction	☐ Not reported	☐ Reported (Give value:)	
	Equivalence (item-objective internal consistency)	☐ Not reported	☐ Reported (Give value:)	
	Stability	☐ Not reported	☐ Reported (Give value:)	
	Cultural Bias	☐ Not reported	☐ Reported (Give value:)	
4. TEST QUALITY (judgmental data)	*Test-Objective Congruence (content validity)*	0(0) ☐ *Not reported*	1(1) ☐ *Reported (Give value:)*	
	Cultural Bias	0(0) ☐ *Not reported*	1(1) ☐ *Reported (Give value:)*	
5. TEST QUALITY (response data)	*Sensitivity to Instruction*	0(0) ☐ *Not reported*	1(2) ☐ *Reported (Give value:)*	
	Equivalence (internal consistency by objective)	0(0) ☐ *Not reported*	2(2) ☐ *Reported (Give value:)*	
	Stability (test-retest/ alternate forms)	0(0) ☐ *Not reported*	2(2) ☐ *Reported (Give value:)*	
	Stability (number of items per objective)	0(0) ☐ *Not reported*	2(2) ☐ *Reported (Give value:)*	
	Criterion Validity	0(0) ☐ *Not reported*	2(2) ☐ *Reported (Give value:)*	
	Cultural Bias	0(0) ☐ *Not reported*	2(2) ☐ *Reported (Give value:)*	

1: Marketing and Packaging*				
1. SCOPE (of Total CRT)	Grade Levels Tested (total CRT) K 1 2 3 4 5 6 7 8 9 10 11 12			
	Grade Level Coverage (total CRT)	0(0) ☐ *Doesn't cover needed grades*	0(1) ☐ *Covers needed grades*	
2. TEST ORGANIZATION	*Number of Separate Tests*	2(0) ☐ *1 test*	0(1) ☐ *2-9 tests*	0(2) ☐ *10 or more tests* ☐ *Not applicable. No pre-formatted tests*
	List of Objectives	☐ No	☐ Yes	
	Flexibility to Select Items	☐ No	☐ Yes, some	☐ Yes, extensive
	Answer Sheet Format	☐ Hand scoreable only	☐ Machine scoreable only	☐ Hand or machine scoreable
	Test Length (range)	☐ ___ to ___ objectives per test	☐ ___ to ___ items per objective	
	Alternate Forms	☐ No	☐ Yes	
3. AVAILABLE MATERIALS	Tests	☐ No	☐ Yes, standard	☐ Yes, optional
	Technical Manual	☐ No	☐ Yes, standard	☐ Yes, optional
	User's Manual/Guides	☐ No	☐ Yes, standard	☐ Yes, optional
	Answer Sheets	☐ No	☐ Yes, standard	☐ Yes, optional
	Cassettes/Special Equipment	☐ No	☐ Yes, standard	☐ Yes, optional
	Student Report Forms	☐ No	☐ Yes, standard	☐ Yes, optional
	Resource Books	☐ No	☐ Yes, standard	☐ Yes, optional
4. OTHER PUBLISHER SERVICES	Inservice Training Available	☐ No	☐ Yes, standard	☐ Yes, optional
	Scoring of Tests	☐ No	☐ Simple scores only	☐ Extensive score summaries
5. COSTS	Cost per Student at a Given Grade Level			
6. QUALITY OF MATERIALS	Physical Quality of Materials	☐ Poor	☐ Good	☐ Very good

*Information sources for each component of the CRTDE may be found in the CRT's test booklets, examiner's manual, and/or technical manual. It should be noted that not all CRTs have these three parts nor is information similarly organized from publisher to publisher.

RATING INSTRUCTIONS FOR THE CRTDE

1. MARKETING AND PACKAGING

1. SCOPE (Scope of the total CRT system; that is, taking into account all tests for all grade or achievement levels considered together.)
   a) *Grade Levels Tested* (total CRT)
      K-12—Circle each grade level for which test materials are available.
   b) *Grade Level Coverage* (total CRT)*
      0 (0) Doesn't cover needed grades—Forms of the test are not available for all needed grade levels.

      2 (1) Covers needed grades—Forms of the test are available for each needed grade level.

2. TEST ORGANIZATION**
   a) *Number of Separate Tests*
      (Most CRTs are organized into a series of independent tests each measuring one or a few objectives; some have just one test per grade level and others have no preset tests, simply an item pool from which tests are made to order.)

      2 (0) 1 test—Only one preset test is provided at this achievement/grade level.

      0 (1) 2-9 tests—Two to nine preset tests are provided at this achievement/grade level.

      0 (2) 10+ tests—Ten or more preset tests are provided at this achievement/grade level.

      0 (0) Not applicable—No preset tests are provided.
   b) *List of Objectives*
      No—A list of the CRT objectives is not provided.

      Yes—A list of the CRT objectives is provided.
   c) *Flexibility to Select Items*
      No—All items for a given objective must be used. This is usually the case with pre-formatted tests.

      Yes, some—There is some opportunity to select the items that can be used to test an objective. This is the case with CRTs that provide detailed item-writing rules, since parallel items can be generated.

      Yes, extensive—There is freedom to choose items to test an objective. An example of this situation is a CRT that has an item pool from which tests are custom-made.
   d) *Answer Sheet Format*
      Hand scoreable only—Tests must be scored manually.

      Machine scoreable only—Tests must be machine scored.

      Hand or machine scoreable—Tests can be scored manually and by machine.
   e) *Test Length* (range)
      ____ to ____ objectives per test—Range in number of objectives per test

      ____ to ____ items per objective—Range in the number of items per objective
   f) *Alternate Forms* (Parallel tests that measure the same content using different but equivalent test items)

      No—Alternate test forms are not provided.

      Yes—Alternate test forms are provided.

3. AVAILABLE MATERIALS. (The kinds of materials that are provided as part of the CRT. For each item in this subsection, a distinction is made between those materials that are provided as standard parts of the CRT and the materials that are optional [that can be purchased separately].)
   a) *Tests* (Pre-formatted test forms)
      No—Not provided as part of the CRT

      Yes, standard—Provided as a regular part of the CRT

      Yes, optional—Not automatically included as part of the CRT; can be purchased separately
   b) *Technical Manual* (This is a report that describes the CRT system and the way in which it was field tested and analyzed. This manual should present statistical indexes of reliability and validity.)

      No—Not provided as part of the CRT

      Yes, standard—Provided as a regular part of the CRT

      Yes, optional—Not automatically included as part of the CRT; can be purchased separately
   c) *User's Manual/Guides* (A manual or brochure written for teachers and others who will administer the CRTs, the manuals usually include detailed instructions for test administration and use of the system.)

      No—Not provided as part of the CRT

      Yes, standard—Provided as a regular part of the CRT

      Yes, optional—Not automatically included as part of the CRT; can be purchased separately
   d) *Answer Sheets* (Pre-formatted forms for recording answers)

      No—Not provided as part of the CRT

      Yes, standard—Provided as a regular part of the CRT

      Yes, optional—Not automatically included as part of the CRT; can be purchased separately.
   e) *Cassettes/Special Equipment* (Tape cassettes, video equipment, etc., required to administer and/or score the CRT)

      No—Not provided as part of the CRT

      Yes, standard—Provided as a regular part of the CRT

      Yes, optional—Not automatically included as part of the CRT; can be purchased separately

*Assign weights to all items in italics. If the CRT is going to be used in an evaluation context, use the weights outside the parentheses; if it is going to be used in a classroom context, use the weights in the parentheses. Record the value of the weight in the box.

**Note: From this point forward, it is assumed that a CRT for a single grade or achievement level is being reviewed.

f) *Student Report Forms* (Special forms for documenting the progress of each student)

No—Not provided as part of the CRT

Yes, standard—Provided as a regular part of the CRT

Yes, optional—Not automatically included as part of the CRT; can be purchased separately

g) *Resource Books* (Special aids associated with the CRT—for example, curriculum guides that reference pages in text where the CRT's objectives are covered)

No—Not provided as part of the CRT

Yes, standard—Provided as a regular part of the CRT

Yes, optional—Not automatically included as part of the CRT; can be purchased separately

4. OTHER PUBLISHER SERVICES (Adjunct services available from the publisher)

a) *Inservice Training Available* (Instruction in the use of the CRT system)

No—Publisher does not make available any form of inservice training.

Yes, standard—Publisher offers inservice training as a standard part of the CRT.

Yes, optional—Publisher offers inservice training which can be purchased separately from the CRT.

b) *Scoring of Tests* (Scoring services that usually involve sending tests to the publisher for scoring or the rental/purchase of special computer programs for scoring tests)

No—Scoring services are not available from the publisher.

Simple scores only—Only individual test-scoring services are available (no summary information).

Extensive score summaries—In addition to individual test scores, aggregated scores and other summary data are available.

5. COSTS (Of purchasing the CRT system from the publisher)

Cost per Student at a Given Grade Level. (Cost per student of the CRT at a given grade or achievement level is based on a minimum purchase. Thus, if the test is sold in lots of 35, the cost per student would be 1/35 of the lot price.)

6. QUALITY OF MATERIALS

Physical Quality of the Materials (A subjective evaluation of the CRT materials)

Poor—Below average quality

Good—Average quality

Very good—Above average quality

2. EXAMINEE APPROPRIATENESS

1. TEST ITEMS

a) *Study of Test Items' Appropriateness* (This includes investigations, conducted by the publisher, for the appropriateness of the test items for the intended examinees. Such investigations are usually documented in technical manuals or user's guides.)

Not reported—An investigation of test items' appropriateness is not reported. (Note: mention that an investigation was conducted without any details should be rated "not reported." Some details on the nature of the investigation and/or its results are required.)

Expert judgment—Experts' opinions are used to establish test items' appropriateness.

Response data—Empirical studies (that include giving test items to examinees) were conducted in order to establish test items' appropriateness.

b) *Vocabulary, Brevity, Clarity* (A subjective evaluation of test items' appropriateness in terms of their vocabulary, brevity, and clarity)

Inappropriate—Most items are inappropriate. That is, the items use vocabulary that is too difficult or easy; the items are needlessly long; the items are misleading; or there is no connection between the item stems and answers.

Appropriate—Most items are appropriate. That is, the vocabulary used matches the intended examinees' educational level; the items are not too long and contain only relevant information; and there is a simple connection between the item stems and answers.

c) *Tasks Required of Examinees* (A subjective evaluation of the ability of the designated examinees to accomplish tasks required to complete the test items.)

Inappropriate—Most items involve tasks that are too easy or difficult for examinees.

Appropriate—Most items involve tasks that examinees should be able to accomplish.

2. INSTRUCTIONS

a) *Vocabulary, Brevity, Clarity* (A subjective evaluation of the instructions, either read to or by examinees, in terms of their vocabulary, brevity, and clarity)

Inappropriate—Instructions are usually inappropriate. That is, the vocabulary used is too easy or too difficult for examinees, they are needlessly long, or they are misleading and confusing.

Appropriate—Instructions are usually appropriate. That is, the vocabulary used matches examinees' educational level; they are brief and easily understood.

b) *Illustrative Sample Items* (The inclusion of sample test items in the instructions)

Not present—No sample items provided.

Present but not clarifying—Sample items are provided, but are not representative of the items.

Effective and clarifying—Sample items are provided that accurately represent the tasks required by the test items.

2. Examinee Appropriateness				
1. TEST ITEMS	Study of Test Item's Appropriateness Vocabulary, Brevity, Clarity Tasks Required of Examinees	☐ Not reported ☐ Inappropriate ☐ Inappropriate	☐ Expert judgment	☐ Response data ☐ Appropriate ☐ Appropriate
2. INSTRUCTIONS	Vocabulary, Brevity, Clarity Illustrative Sample Items	☐ Inappropriate ☐ Not present	☐ Present but not clarifying	☐ Appropriate ☐ Effective and clarifying
3. FORMAT	Test Page Layout Illustrations and Print Auditory Presentation	☐ Complicated ☐ Unclear ☐ Garbled		☐ Clear ☐ Clear ☐ Clear
4. TIMING	Timing and Pacing	☐ Inappropriate	☐ Appropriate	☐ No guidelines
5. RECORDING ANSWERS	Response Scheme	☐ Complicated		☐ Simple

3. FORMAT (Appropriateness of the formatting of the CRT materials)

a) *Test Page Layout* (A subjective evaluation of the arrangement of written materials on a test page)

Complicated—Below average quality; crowded and confusing format

Clear—Average or above average quality; clear format

b) *Illustrations and Print* (A subjective evaluation of the clarity of print and illustrations)

Unclear—Below average quality; difficult to follow and confusing

Clear—Average or above average quality; readable; realistic; up-to-date, and bold

c) *Auditory Presentation* (A subjective evaluation of the clarity and ease of understanding of oral presentations)

Garbled—Below average quality; garbled presentation using slang and/or poorly paced

Clear—Average or above average quality; easily understood presentation with no slang and well-paced

4. TIMING

Timing and Pacing (A subjective evaluation of the timing guidelines)

Inappropriate—Suggested timing and pacing techniques are usually inappropriate for the designated examinees' educational level or amount of time available for testing.

Appropriate—Suggested timing and pacing techniques are usually appropriate for examinees.

No guidelines—No timing and pacing guidelines are given.

5. RECORDING ANSWERS

Response Scheme

Complicated—The procedure used to record answers to the test items is difficult to use and likely to be confusing to examinees.

Simple—The procedures used to record answers are simple and easy to use (multiple choice or fill-ins).

3. ADMINISTRATIVE USABILITY

1. ADMINISTRATION

a) *Size of Testing Group*

0 (1) Individual only—Test(s) *must* be administered on an individual basis.

2 (1) Groups only— Test(s) may be administered to small groups (fewer than 30).

2 (2) Individual or group—Test(s) may be administered to large groups (more than 30). Group administration includes a single cassette that can be used by many students simultaneously.

b) *Administrator*

0 (0) Specialist—Only a specialist (such as a psychologist) may administer the CRT.

1 (1) School personnel—Teachers or classroom aides may administer the CRT.

1 (2) Self-administration—The CRT can be administered without assistance from teachers or others. This includes CRT systems with audio-cassette test directions.

c) *Administration Time*

Shortest possible time—Shortest time recommended for examinees to complete any single test

Longest possible time—Longest time recommended for examinees to complete any single test.

No time limits—No limits on testing time are provided.

d) *Directions*

Not available—There are no directions to be read to or by the examinee.

Available, incomplete—Directions do not cover all aspects of test taking; a standardized system for test taking is not guaranteed.

Available, complete—Directions cover all aspects of test taking, and a standardized system is established.

e) *Equipment*

Special equipment needed—Special materials other than paper and pencils needed for test administration (such as cassettes and video equipment)

No special equipment needed—Only paper and pencils are needed for test administration.

f) *Order for Testing* (The order in which the objectives measured by the CRT must be tested)

Prescribed sequence—Objectives must be tested in a prescribed order (this is frequently the case, for example, with CRT systems designed for use with a specific curriculum or with CRTs that test all objectives on a single form).

General guidelines—An order for testing objectives is recommended but not mandatory (this is frequently the case when objectives are structured in a hierarchy.

None—There is no prescribed order in which objectives must be tested.

2. ADAPTABILITY (The extent to which the CRT system can be modified by the user)

Modification of CRT by User (Guidelines provided by the publisher for altering or modifying various aspects of the CRT)

No—No modifications are permitted, or no guidelines are given.

Yes, limited—Guidelines are given for making limited changes to the CRT.

Yes, extensive—Guidelines are given for making extensive changes to the CRT.

3: Administrative Usability				
1. ADMINISTRATION	*Size of Testing Group*	0(1) ☐ *Individual only*	2(1) ☐ *Groups only*	2(2) ☐ *Individual or group*
	Administrator	0(0) ☐ *Specialist*	1(1) ☐ *School personnel*	1(2) ☐ *Self-administration*
	Administration Time	☐ Shortest possible time: ____	☐ Longest possible time: ____	☐ No time limits
	Directions	☐ Not available	☐ Available, incomplete	☐ Available, complete
	Equipment	☐ Special equipment needed	☐ No special equipment needed	
	Order for Testing	☐ Prescribed sequence	☐ General guidelines	☐ None
2. ADAPTABILITY	Modification of CRT by User	☐ No	☐ Yes, limited	☐ Yes, extensive
3. SCORING	Ease of Scoring	☐ Subjective	☐ Objective	
4. INTERPRETATION	Score Interpreter	☐ Specialist	☐ School personnel	☐ Self-interpreting
	Score-Interpretation Guide	☐ Not available	☐ Available, complicated	☐ Available, not complicated
	Score Interpretability	☐ Complicated, unusual	☐ Simple, standard	
5. DECISIONS	Decision-Making Utility	☐ No guidelines	☐ Yes, limited	☐ Yes, extensive
	Curriculum Referencing	0(0) ☐ *None*	0(1) ☐ *Yes, some*	0(2) ☐ *Yes, extensive*

3. SCORING

Ease of Scoring

Subjective—Test scores are not assigned using a standardized set of rules and procedures and can be considered a function of scorer's discretion.

Objective—Objective scoring system that is standardized

4. INTERPRETATION

a) *Score Interpreter*

Specialist—A specialist is required to interpret the CRT's scores.

School personnel—Teacher or classroom aides can interpret CRT's scores.

Self-interpreting—Test forms include a scoring mechanism (carbon-backed test form with scoring key and interpretation guide).

b) *Score-Interpretation Guide* (Interpretation guides, in various forms, which permit correct and consistent interpretation of a CRT's scores)

Not available—No guides or directions for interpreting scores are provided.

Available, complicated—Guides are provided to assist in the interpretation of scores but are difficult to understand or have inadequate instructions.

Available, not complicated—Thorough and easy-to-use interpretation guidelines are provided.

c) *Score Interpretability*

Complicated, unusual—Interpretation systems are not commonly used and/or require numerous tables and mathematical conversions.

Simple, standard—Interpretation systems are generally understood and easily used.

5. DECISIONS

a) *Decision-Making Utility* (The usefulness of guidelines provided by the publisher that describe how to use the CRT results to make educational decisions)

No guidelines—Guidelines for rules for decision making are not provided.

Yes, limited—Guidelines are provided, but they are vague, incomplete, or not particularly relevant to the test's stated purposes.

Yes, extensive—Guidelines are provided that are complete, clearly defined, and relevant to the test's stated purposes.

b) *Curriculum Referencing* (A guide, usually organized by objective, linking each of the CRT's objectives to specific components of major instructional programs, is provided.)

0 (0) None—The publisher provides no system of curriculum referencing.

0 (1) Yes, some—The publisher provides a referencing system that is limited in that not all the CRT's objectives are referenced and/or only a small number of instructional programs are included.

0 (2) Yes, extensive—The publisher provides a referencing system that includes all the CRT's objectives and most of the major instructional programs.

4. FUNCTION AND PURPOSE

1. PURPOSE (The purpose of the CRT suggested by the publisher)

Test Uses

0 (1) Diagnosis—The CRT is used to identify difficulties with specific learning objectives, tasks, and/or behaviors.

0 (1) Placement—The CRT is used to locate the examinee's position in a curriculum or learning hierarchy.

1 (2) Achievement/progress—The CRT is used to measure achievement of specific learning objectives, tasks, and/or behaviors.

2 (1) Achievement/outcomes—The CRT is used to measure the outcomes of instruction and/or the extent to which an educational program's objectives have been achieved.

2 (0) Comparison of instructional programs—The CRT is used to compare two or more educational programs.

Other, please specify—Name other uses suggested for CRT.

5. OBJECTIVES DEVELOPMENT

1. DOMAIN (OBJECTIVES) SPECIFICATION

a) *Domain Definition* (How the domain or the organized set of objectives measured by the CRT is defined)

None—Domain is not reported and/or defined.

Vague, general—Domain is defined in unclear or in very general terms.

Specific—Domain is defined clearly and in detail.

b) *Domain Structure*

Content area—Domain structure is not clearly specified (for example, the domain is only linked to a broad content area).

Content/process matrix—Domain is structured by a content/process matrix that defines the knowledge that will be assessed and the ways in which it will be measured.

Objectives/item generation format—Formal, replicable rules are given for generation of items and/or test items.

2. OBJECTIVES' CHARACTERISTICS

a) *Organization of Objectives*

None—A complete list of objectives that define the CRT for the grade/level being reviewed is not provided.

4: Function and Purpose				
1. PURPOSE	*Test Uses (Circle all that apply)*	0(1) ☐ *Diagnosis* 2(1) ☐ *Achievement/outcomes* ☐ Other, please specify:	0(1) ☐ *Placement* 2(0) ☐ *Comparison of instruction programs*	1(2) ☐ *Achievement/progress*

5: Objectives Development					
1. DOMAIN (OBJECTIVE) SPECIFICATION	Domain Definition Domain Structure	☐ None ☐ Content area	☐ Vague, general ☐ Content/process matrix	☐ Specific ☐ Objectives/item generation format	
2. OBJECTIVES' CHARACTERISTICS	Organization of Objectives Level of Generality of Objectives	☐ None ☐ General	☐ Simple list (no structure) ☐ Specific	☐ Categories (strands) ☐ Very detailed	☐ Hierarchy
3. MATCH TO INSTRUCTION	*Curriculum Match*	2(0) ☐ *None*	1(1) ☐ *Some*	0(2) ☐ Extensive	

Simple list (no structure)—A list of objectives that define the CRT is given, but the objectives are not structured or organized in any specific fashion.

Categories (strands)—A list of the objectives defining the CRT is provided and the objectives are organized into major skill areas or strands (in reading, two strands that might be used to organize the objectives are comprehension and vocabulary).

Hierarchy—A list of the objectives that define the CRT is provided with the objectives organized within categories into a hierarchy of skills/tasks.

b) *Level of Generality of Objectives* (How broadly or narrowly objectives are stated)
General—Very global statements cover a wide range of content, skills, and behavior.

Specific—Statements clearly define the skill or knowledge being assessed but are not so specific as to constitute behavioral objectives.

Very detailed—Objectives are stated in detail or in behavioral terms.

3. MATCH TO INSTRUCTION
Curriculum Match
2 (0) None—The CRT system is not designed for use with a specific instructional program.

1 (1) Some—The CRT system is not necessarily dependent on the skills or context of an instructional program. However, it may be more appropriately used with certain types of programs (for example, a CRT may be developed from several instructional programs and reflect the bias of these programs, or the CRT might emphasize terminology and nomenclature used in only some programs).

0 (2) Extensive—The CRT, its objectives, and test items are dependent on a particular curriculum or set of instructional materials and techniques.

6. ITEM DEVELOPMENT

1. ITEM-OBJECTIVE RELATION
a) *Items Coded to Objectives*
0 (0) No—Items are not referenced to a specific objective(s).

2 (2) Yes—Each test item is referenced to a specific objective(s).

b) *Number of Items per Objective* (The minimum, maximum, and average number of items used to test each objective)

c) *Scope of Coverage of Objectives*
Poor—In general, test items do not adequately cover the range of behaviors, contents, situations, and/or skills that are associated with the objectives being tested.

Good—Most test items adequately cover the range of skills, behaviors, contents, and/or situations associated with the objective being tested.

2. ITEM GENERATION
Rules for Item Writing (A procedure or set of rules for writing test items)
None—No system/rules were used (or reported) to guide item writing.

Suggestions—Some very general rules were provided (and reported) to guide item writing (all items must be multiple choice, for example).

Specifications—Comprehensive, detailed system/rules were provided (and reported) to guide item writing. Such rules should limit the kinds of items used to measure an objective (define appropriate content and format, for example).

7. METHODS OF SCORE INTERPRETATION

1. CRITERION-REFERENCED
Scores Reported in Terms of Level of Performance (Criterion-referenced test scores for individuals and groups must be presented in terms of the level of competency or mastery of the specific objectives on which the CRT is based. The distinctive feature of a CRT score must, therefore, lie in its emphasis on describing the absolute rather than the relative level of performance with respect to an objective or skill.) Some of the different kinds of CRT scores include:
0 (0) Item difficulty—This represents the percentage of examinees or groups who "pass" each item; that is, the item's difficulty.

0 (0) Actual score—This is the number or percent of correct items on a given objective, referring to the number of items actually passed on the test.

1 (1) True score—This indicates an individual's or group's true level of performance on an objective, referring to the portion of the total universe of items for an objective that an individual or group could answer correctly. (That is, if every possible item was tested, this score is the number of items that an individual or group would pass.)

0 (0) Arbitrary mastery—This refers to whether an individual or group has achieved a pre-set but arbitrarily defined level of performance.

1 (1) Empirical mastery—This refers to whether an individual or group has achieved a pre-set criterion level of performance where the criterion level is educationally meaningful and empirically justified.

1 (1) Achievement level after a certain amount of learning time—This reports the time it takes (in class hours or calendar days) for an examinee or group to achieve a given performance level.

1 (1) Probability of achieving next level—This refers to the probability that the examinee is ready to begin the next level of instruction (this may be based on both the number of items correct and the patterns of answers given to these items).

6: Item Development				
1. ITEM-OBJECTIVE RELATION	*Items Coded to Objectives*	0(0) ☐ *No*	2(2) ☐ *Yes*	
	Number of Items per Objective	☐ Not applicable–Items not coded to objectives	☐ Range: ___ to ___ items/objective	☐ Average: ___
	Scope of Coverage of Objectives	☐ Poor	☐ Good	
2. ITEM GENERATION	Rules for Item Writing	☐ None	☐ Suggestions	☐ Specifications

7: Methods of Score Interpretation				
1. CRITERION-REFERENCED	*Scores Reported in Terms of Level of Performance (Check all that apply)*	0(0) ☐ *Item difficulty* 0(0) ☐ *Arbitrary mastery* 1(1) ☐ Achievement level after a certain amount of learning time	0(0) ☐ *Actual score (Percent of items correct)* 1(1) ☐ *Empirical mastery* 1(1) ☐ Probability of achieving next level	1(1) ☐ *True score (Percent of objective achieved)*
2. NORM-REFERENCED	*Comparative Scores Reported as well as a Criterion-Referenced Score (Check all that apply)*	1(1) ☐ *Standard score scales*	1(1) ☐ *Age/grade equivalents*	1(1) ☐ *Percentiles*

27

2. NORM-REFERENCED

Comparative Scores Reported as Well as a Criterion-Referenced Score (Individual's or groups' scores must be interpreted in relation to the scores of other individuals or groups who have taken or who might take the test.)

1 (1) Standard score scales and age/grade equivalents—These describe an individual's or group's expected performance at given grade levels.

1 (1) Percentiles—These describe scores in terms of the ranking or percentage of individuals whose scores fall below a given score.

8. ANALYSIS AND VALIDATION

1. FIELD TEST (A field test of the published version of the CRT system in which examinee response data are used to establish the reliability and validity of the items and test system. Field tests should not be confused with pilot tests of unpublished, preliminary working drafts of the CRT system.)

a) *Field Test Reported*

No—Field test was not conducted, or field test was not documented.

Yes—Field-test methods and results are reported with some detail.

b) *Scale*

Small—Field test involved just one or two schools or a single school district.

Moderate—Field test involved several school districts.

Large—Field test involved students from many school districts.

c) *Scope-Geographic*

Local—Field test was restricted to one city or county.

Regional—Field test involved a specific region of the country.

National—Field test sites are geographically representative of the nation.

d) *Scope-Ethnic*

Little ethnic representation—Minority groups are not included in sufficient numbers in field test (cannot measure with confidence the relative performance of minority groups).

Ethnic representation—Minority groups are included in sufficient numbers in field test.

e) *Sample Representativeness*

No sampling plan—Participants in field test were selected without any predetermined sampling plan (schools volunteered for field testing).

Probability sampling plan—Participants in the field test were selected using random sampling or random stratified sampling.

Special sampling (non-probability)—Participants in field test were selected using a systematic but non-probabilistic plan.

2. ITEM QUALITY/JUDGMENTAL DATA*

a) *Item-Objective Congruence* (The extent to which an item measures the relevant objective)

Not reported—No judgmental data on item-objective congruence are reported.

Reported (Give value)—Some judgmental data on item-objective congruence are reported.

b) *Cultural Bias* (The existence of systematic differences in performance on an item across different cultural groups)

Not reported—No judgmental data on item-objective congruence are reported.

Reported (Give value)—Some judgmental data on item-objective congruence are reported.

3. ITEM QUALITY/RESPONSE DATA

a) *Sensitivity to Instruction* (An item's ability to discriminate between those who have and have not benefited from instruction)

Not reported—No estimate of sensitivity to instruction based on response data is reported.

Reported (Give value)—Some estimate of sensitivity to instruction is reported.

b) *Equivalence* (The internal consistency for an item associated with a particular objective; the extent to which a test item behaves similarly to other items measuring the same objective)

Not reported—No estimate of sensitivity to instruction based on response data is reported.

Reported (Give value)—Some estimate of sensitivity to instruction is reported.

c) *Stability* (The extent to which performance on an item remains constant over time)

Not reported—no estimate of sensitivity to instruction based on response data is reported.

Reported (Give value)—Some estimate of sensitivity to instruction is reported.

d) *Cultural Bias*

Not reported—No estimate of sensitivity to instruction based on response data is reported.

Reported (Give value)—Some estimate of sensitivity to instruction is reported.

4. TEST QUALITY/JUDGMENTAL DATA

a) *Test Objective Congruence* (The extent to which a test measures the relevant objective; content validity)

0 (0) Not reported—No judgmental data on item-objective congruence are reported.

1 (1) Reported (Give value)—Some judgmental data on item-objective congruence are reported.

b) *Cultural Bias* (The existence of systematic differences in performance on the test across cultural groups)

*Different methods can be used to determine the quality of individual test items and the quality of the total test. A distinction is made according to the kinds of data (judgmental and response) used to determine item and test quality. Judgmental data refer to reviews of the test materials by experts and other persons who might use the system. Reponse data refer to the use of participants' scores from field tests of the CRT materials.

8. Analysis and Validation				
1. FIELD TEST	Field Test Reported	☐ No	☐ Yes	
	Scale	☐ Small	☐ Moderate	☐ Large
	Scope-Geographic	☐ Local	☐ Regional	☐ National
	Scope-Ethnic	☐ Little ethnic representation	☐ Ethnic representation	
	Sample Representativeness	☐ No sampling plan	☐ Probability sampling plan	☐ Special sampling (non-probability)
2. ITEM QUALITY (judgmental data)	Item-Objective Congruence	☐ Not reported	☐ Reported (Give value:)	
	Cultural Bias	☐ Not reported	☐ Reported (Give value:)	
3. ITEM QUALITY (response data)	Sensitivity to Instruction	☐ Not reported	☐ Reported (Give value:)	
	Equivalence (item-objective internal consistency)	☐ Not reported	☐ Reported (Give value:)	
	Stability	☐ Not reported	☐ Reported (Give value:)	
	Cultural Bias	☐ Not reported	☐ Reported (Give value:)	
4. TEST QUALITY (judgmental data)	*Test-Objective Congruence (content validity)*	0(0) ☐ *Not reported*	1(1) ☐ *Reported (Give value:)*	
	Cultural Bias	0(0) ☐ *Not reported*	1(1) ☐ *Reported (Give value:)*	
5. TEST QUALITY (response data)	*Sensitivity to Instruction*	0(0) ☐ *Not reported*	1(2) ☐ *Reported (Give value:)*	
	Equivalence (internal consistency by objective)	0(0) ☐ *Not reported*	2(2) ☐ *Reported (Give value:)*	
	Stability (test-retest/ alternate forms)	0(0) ☐ *Not reported*	2(2) ☐ *Reported (Give value:)*	
	Stability (number of items per objective)	0(0) ☐ *Not reported*	2(2) ☐ *Reported (Give value:)*	
	Criterion Validity	0(0) ☐ *Not reported*	2(2) ☐ *Reported (Give value:)*	
	Cultural Bias	0(0) ☐ *Not reported*	2(2) ☐ *Reported (Give value:)*	

0 (0) Not reported—No judgmental data on item-objective congruence are reported.

1 (1) Reported (Give value)—Some judgmental data on item-objective congruence are reported.

5. TEST QUALITY/RESPONSE DATA

a) *Sensitivity to Instruction* (A test's ability to discriminate between those who have and those who have not benefited from instruction)

0 (0) Not reported—No estimate of sensitivity to instruction based on response data is reported.

1 (2) Reported (Give value)—Some estimate of sensitivity to instruction is reported.

b) *Equivalence* (Internal consistency, the extent to which all items that measure a given objective behave similarly)

0 (0) Not reported—No estimate of sensitivity to instruction based on response data is reported.

2 (2) Reported (Give value)—Some estimate of sensitivity to instruction is reported.

c) *Stability* (Test-retest; alternate forms; the extent to which test performance remains constant over time)

0 (0) Not reported—No estimate of sensitivity to instruction based on response data is reported.

2 (2) Reported (Give value)—Some estimate of sensitivity to instruction is reported.

d) *Stability* (Number of items per objective; a determination of the number of items needed to obtain a stable score on an objective)

0 (0) Not reported—No estimate of sensitivity to instruction based on response data is reported.

2 (2) Reported (Give value)—Some estimate of sensitivity to instruction is reported.

e) *Criterion Validity* (A determination of the criterion in terms of which CRT scores are reported)

0 (0) Not reported—No estimate of sensitivity to instruction based on response data is reported.

2 (2) Reported (Give value)—Some estimate of sensitivity to instruction is reported.

f) *Cultural Bias* (The existence of systematic differences in test performance across cultural groups; this can be measured by regression techniques)

0 (0) Not reported—No estimate of sensitivity to instruction based on response data is reported.

2 (2) Reported (Give value)—Some estimate of sensitivity to instruction is reported.

REFERENCES*

1. Anderson, S., *et al. Encyclopedia of educational evaluation.* San Francisco: Jossey-Bass, 1975.
2. Baker, E.L. Using measurement to improve instruction. Paper presented at Convention of American Psychological Association, Honolulu, Hawaii, 1972. ED 069 762.
3. Baker, R.L. Measurement considerations in instruction product development. Paper presented at Conference on Problems in Objectives-Based Measurement, Center for the Study of Evaluation, Univer. of California, 1972.
4. Cronbach, L.J. *Essentials of psychological testing.* (3rd ed.) New York: Harper, 1970.
5. Cronbach, L.J. Test validation. In L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.
6. Ebel, R.L. Evaluation and educational objectives: behavioral and otherwise. Paper presented at the Convention of the American Psychological Association, Honolulu, Hawaii, 1972.
7. Glaser, R., & Nitko, A. Measurement in learning and instruction. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971. Pp. 652-670.
8. Harris, C. Comments on problems of objectives-based measurement. Paper presented at the annual meeting, American Educational Research Association, New Orleans, 1973.
9. Harris, M.L., and Stewart, D.M. Application of classical strategies to criterion-referenced test construction. Paper presented at the annual meeting, American Educational Research Association, New York, 1971.
10. Hively, W. Introduction to domain-referenced achievement testing. Symposium presentation, American Educational Research Association, Minnesota, 1970.
11. Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. Domain-referenced curriculum evaluation: a technical handbook and a case study from the MINNEMAST project. CSE Monograph Series in Evaluation, Volume 1. Center for the Study of Evaluation, Univer. of California, Los Angeles, 1973.
12. Hoepfner, R., Conniff, W. Jr., Petrosko, J., Watkins, J., Erlich, O., Todaro, R., and Hoyt, M. *CSE secondary school test evaluations: grades 7 and 8.* Center for the Study of Evaluation, Univer. of California, Los Angeles, 1974. ED 113 382.
13. Keesling, James W. Identification of differing intended outcomes and their implications for evaluation. Paper presented at the annual meeting, American Educational Research Association, Washington, D.C., 1975.
14. Klein, S.P. Evaluating tests in terms of the information they provide. *Evaluation Comment,* 1970, 2, No. 2, 1-6. ED 045 699.
15. Klein, S.P. An evaluation of New Mexico's educational priorities. Paper presented at Western Psychological Association, Portland, Oregon, 1972. ED 077 938.
16. Mager, R.F. *Preparing instructional objectives.* San Francisco: Fearon Publishers, Inc., 1962.
17. Novick, M.R. and Lewis, C. Prescribing test length for criterion-referenced measurement. CSE *Monograph No. 3.* Center for the Study of Evaluation, Univer. of California, Los Angeles, 1974.
18. Popham, W., & Husek, T.R. Implications of criterion-referenced measurement. *Journal of Educational Measurement,* 1969, 6, No. 1, 1-9.
19. Popham, W.J. *Educational evaluation.* Englewood Cliffs, N.J.: Prentice-Hall, 1975.
20. Skager, R. Generating criterion-referenced tests from objectives based assessment systems: unsolved problems in test development, assembly and interpretation. Paper presented at the annual meeting, American Educational Research Association, New Orleans, 1973.
21. Skager, R. Critical differentiating characteristics for tests of educational achievement. Paper presented at the annual meeting, American Educational Research Association, Washington, D.C. 1975.
22. Wilson, H.A. A humanistic approach to criterion-referenced testing. Paper presented at the annual meeting, American Educational Research Association, New Orleans, 1973. ED 081 842.

*Items followed by an ED number (for example, ED 099 429) are available from the ERIC Document Reproduction Service (EDRS). Consult the most recent issue of *Resources in Education* for the address and ordering information.